

Environmental Aerosol Sample Analysis

ACHLA MARATHE ¹ SHERRY PARKER ² JANE BOOKER ²

1 Introduction

This project aims at developing a real time biological agent detection system which can detect the presence of living biological agents in the environment. If developed successfully, this system will be a vast improvement over the existing ones which require several hours to do the detection. This system will be capable of collecting, analyzing and characterizing the aerosol samples which will then help in detecting the biological agent attacks.

We first need to develop a model that characterizes the environmental background without any release of biological agents. Once the background and the noise level can be modeled and understood, we can learn about any significant change in the environment due to the release of the biological agents.

2 The MET 1 Data

To measure the baseline conditions, continuous measurements of the aerosol particle size distribution and meteorological information were collected from five different sensors for the period of 28 June to 1 July 1999. These measurements, which are ten seconds apart, bin the particle sizes in six categories, record the wind speed, wind direction and the temperature of the sensor. Before performing the analysis, the data was preprocessed in the following way:

- The raw data has observations every ten seconds and is relatively sparse. So for all the variables, we took averages over every 12 observations or 2 minute time span.
- All the bins were scaled by the following transformation. $\frac{\text{sum}(\text{count}) - \text{count}}{\text{sum}(\text{count})}$. Count is the frequency in each bin. Sum(count) is the total count across 6 bins. This transformation helped in changing the counts to numbers between zero and one. Given that bin3, bin4, bin5 and bin6 often had zero counts in them, models that required ratio of bins could not have been implemented without this transformation.
- Wind speed and direction were scaled to have zero mean and one standard deviation.

¹Los Alamos National Laboratory P.O. Box 1663, MS B265, Los Alamos NM 87545. Email: achla@lanl.gov.

²Los Alamos National Laboratory P.O. Box 1663, MS F600, Los Alamos NM 87545. Email: {sparker, jmb}@lanl.gov.

- The temperature of the sensor was also not included in the analysis because there was very little variation in the temperature over time.
- By doing a separate analysis, we found out that there was significant difference in the observations during day and night time. So to account for that difference, we included a dummy variable for day and night.

$$D_1 = 0 \text{ if day (7am - 7pm)}$$

$$D_1 = 1 \text{ if night (7pm - 7am)}$$

- Similarly, four binary variables were created to reflect the five sensors.
 $S_2 = 1$ if sensor 2
 $S_2 = 0$ otherwise
 $S_3 = 1$ if sensor 3
 $S_3 = 0$ otherwise
 $S_4 = 1$ if sensor 4
 $S_4 = 0$ otherwise
 $S_5 = 1$ if sensor 5
 $S_5 = 0$ otherwise

3 Modeling MET 1 Data

3.1 Methodology

Using the above dataset, several different models and methodologies were formulated and analyzed with the goal of obtaining a stable set of estimated coefficients for the variables described in section 2. We developed a cross validation method, which estimates the coefficients for the training data and as a new observation is given from the test data, it updates the estimated coefficients. If the estimated coefficients show lot of variability due to the additional observation, we get an unstable model. Otherwise the model is stable. The instability of the model can also be measured by the prediction error which is the difference between the actual value of \mathbf{y} and the estimated value of \mathbf{y} . The prediction error is expected to be random and centered around zero. In our model, the response variable $\mathbf{y} = 1$ if there is no biological agent release and $\mathbf{y} = 0$ otherwise. Given this preliminary analysis is being done to just characterize the background, we only have data on $\mathbf{y} = 1$. The traditional statistical time series models did not perform very well due to the lack of information on the response variable \mathbf{y} and excessive variability in the counts as a function of time. The following model was used for the aerosol data analysis.

$\mathbf{d}(v)$ = function of input variable

$$\mathbf{x}(v) = (1, v_1, v_2, v_3, v_1/v_2, v_2/v_3, \dots)^T$$

$$\mathbf{d}(v) = \mathbf{A}^T \mathbf{x}(v)$$

$$\mathbf{y} = (1, 1, 1, 1, \dots)^T$$

Choose \mathbf{A} such that the following sum of squares between \mathbf{y} and its prediction from the model $\mathbf{d}(v)$ is minimized.

$$S_A^2 = E(|\mathbf{d}(v) - \mathbf{y}|^2) = E(|\mathbf{A}^T \mathbf{x}(v) - \mathbf{y}|^2)$$

where E is the expectation operator.

The above optimization problem gives the following adaptive functional structure.

$$A_J = A_{J-1} - \alpha \left(\sum_{j=1}^J x_j x_j^T \right)^{-1} x_J (A_{J-1}^T x_J - y_J)^T$$

where A_J is the next moving set of coefficients based on the previous set, A_{J-1} and $\alpha = (1/\text{number of variables in } \mathbf{x})$ and T is the transpose of a matrix.

$$\left[\sum_{j=1}^J x_j x_j^T \right]^{-1} = \frac{1}{1 - \alpha} \left[\sum_{j=1}^{J-1} x_j x_j^T \right]^{-1} - \alpha \frac{\left[\sum_{j=1}^{J-1} x_j x_j^T \right]^{-1} x_J x_J^T \left[\sum_{j=1}^{J-1} x_j x_j^T \right]^{-1}}{1 + \alpha (x_J^T \left[\sum_{j=1}^{J-1} x_j x_j^T \right]^{-1} x_J - 1)}$$

To obtain the initial guess of the coefficients, \mathbf{A}_0 , we use the training data such that $\mathbf{A}_0 = \left(\sum_{j=1}^J x_j x_j^T \right)^{-1} \sum_{j=1}^J x_j y_j^T$

Once this is calculated, use the test data in blocks of J measurements to update the coefficients and the prediction error which is measured by $(A_{J-1}^T x_J - y_J)^T$.

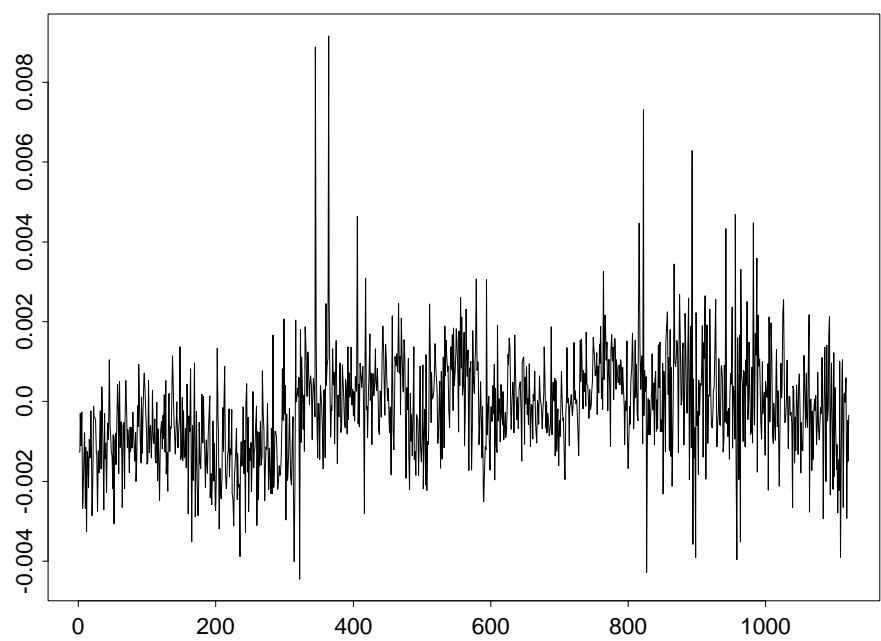
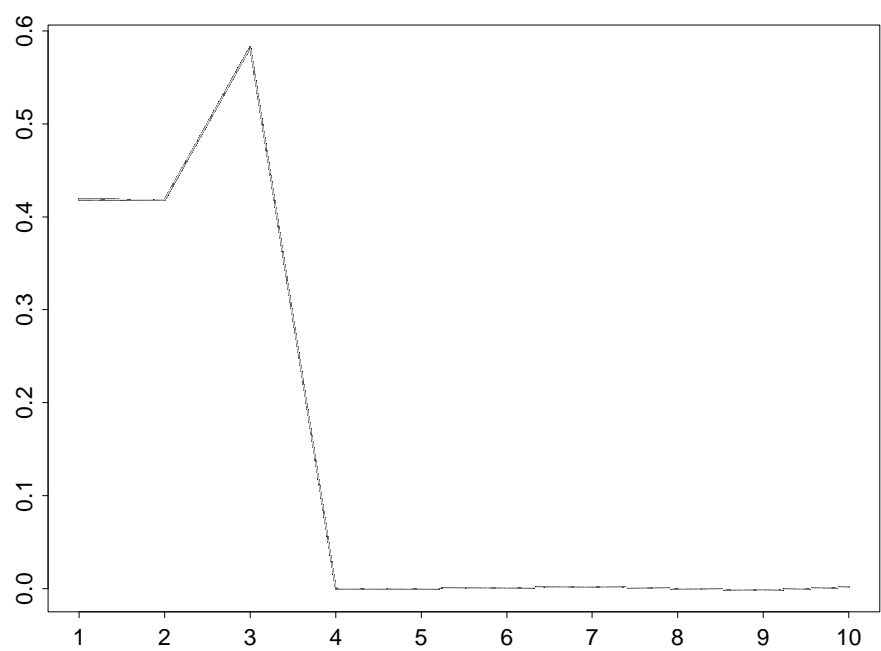
This model was used with different combinations of \mathbf{x} matrices to characterize the background. The parameter estimates of the model for the unperturbed environment will help us later detect the presence of biological agents.

3.2 Results

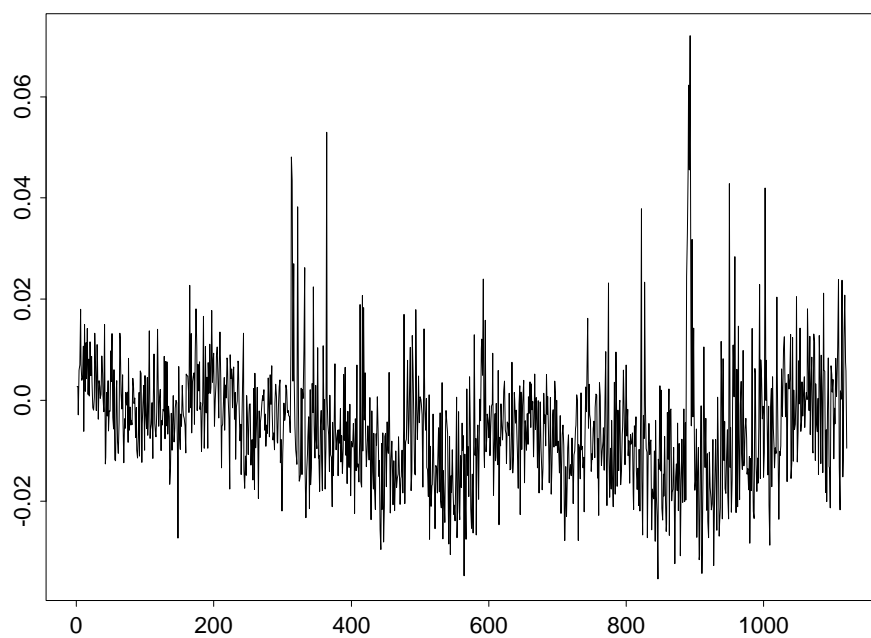
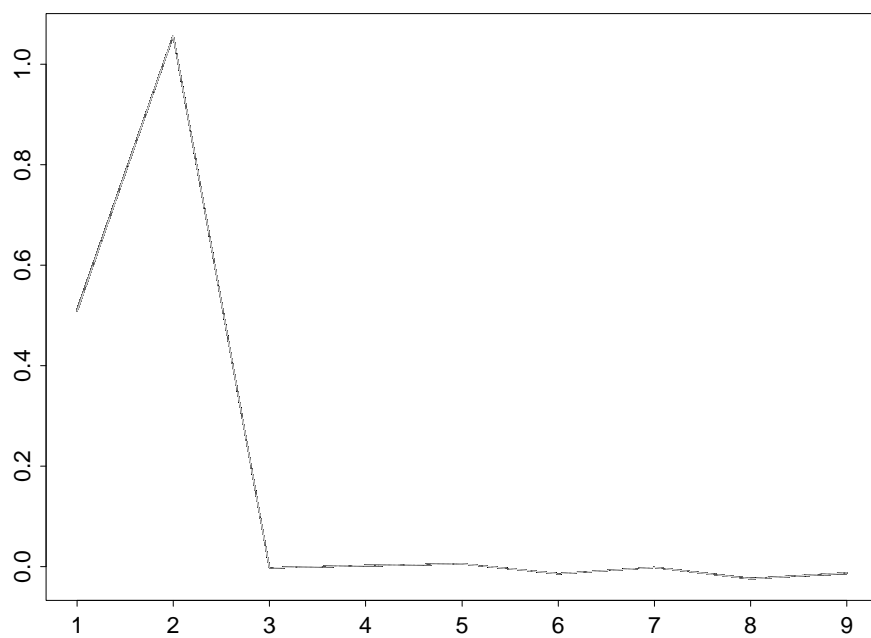
We tested the data using two different training sets. One in which the training data is taken from the first 4 sensor files and the test data is from part of the 4th and all the 5th sensor files. There is no overlap between the training and the test data. The other set had training data from each sensor file and test data also from each sensor file. Again there is no overlap between training and test data. In both cases, we found that the results were similar and stable.

The following models give the desired results of stable coefficients and a random prediction error with zero mean. All the variables have been transformed as explained in section 2 of this report. The training dataset has 5000 data points and the test data 1120 data points.

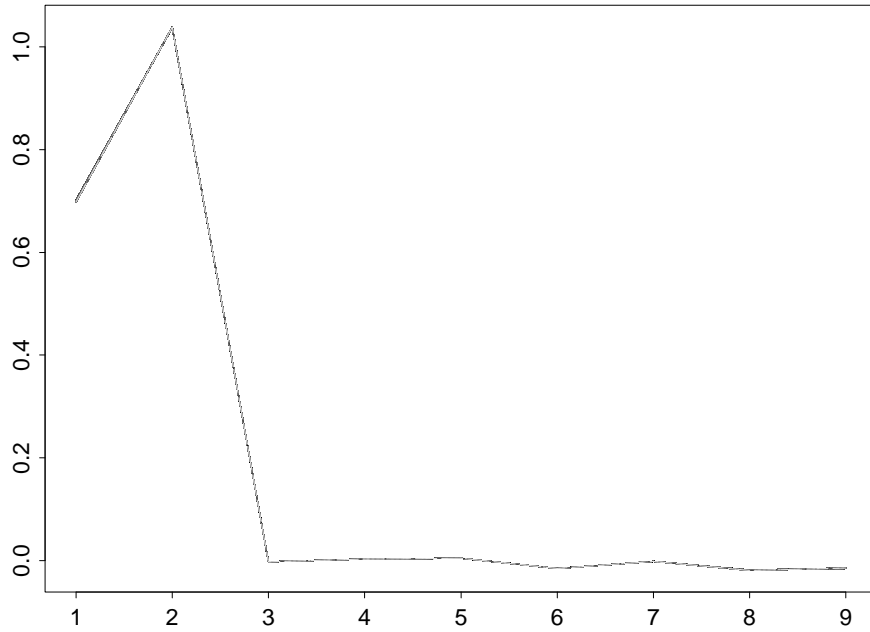
- Model 1: $\mathbf{x}(v) = \{\text{bin1}, \text{bin2}, \text{bin3}, \text{wind speed}, \text{wind direction}, D_1, S_2, S_3, S_4, S_5\}$.

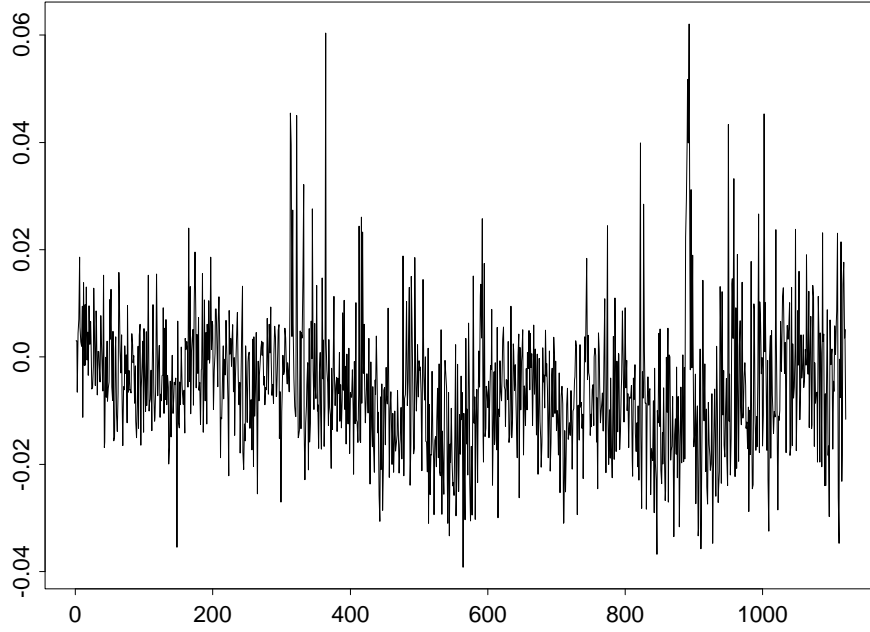


- Model 2: $\mathbf{x}(v) = \{\text{bin1/bin2}, \text{bin2/bin3}, \text{wind speed}, \text{wind direction}, D_1, S_2, S_3, S_4, S_5\}$.



- Given that bin4, bin5 and bin6 had very sparse data, the counts of bin4, bin5, bin6 were collapsed together and added to bin 3. Model 1 and Model 2 were re-estimated using the collapsed bin 3 but no significant change appeared in the values of the coefficients. The reason being that bin 4, bin5 and bin6 have very little information in them. Majority of those values were zero.
- Model 3: $\mathbf{x}(v) = \{\text{bin1}, \text{bin2}, \text{wind speed}, \text{wind direction}, D_1, S_2, S_3, S_4, S_5\}$. Here instead of $\mathbf{y} = 1$, we used bin3 as the response variable.





D_1, S_2, S_3, S_4 and S_5 contribute very little in characterizing the response variable. All of their coefficients are close to zero. However, removing these does increase instability, therefore they remain viable terms in the model. The main determinants are the frequencies in the first three bins which always have non zero coefficients.

3.3 Some Unsuccessful Models

The following models resulted in unstable coefficients and were considered unsuccessful.

- Collapsed counts of bin4, bin5 and bin6 into bin4 and included bin4 in the analysis. $\mathbf{x}(v) = \{\text{bin1}, \text{bin2}, \text{bin3}, \text{bin4}, \text{wind speed}, \text{wind direction}, D_1, S_2, S_3, S_4, S_5\}$.
- Included a constant in the coefficients. $\mathbf{x}(v) = \{\text{constant}, \text{bin1}, \text{bin2}, \text{bin3}, \text{wind speed}, \text{wind direction}, D_1, S_2, S_3, S_4, S_5\}$.
- Ratios of the bins which included bin4, bin5 and bin6 e.g. $\mathbf{x}(v) = \{\text{bin1}/\text{bin2}, \text{bin3}/\text{bin4}, \text{bin5}/\text{bin6}, \text{wind speed}, \text{wind direction}, D_1, S_2, S_3, S_4, S_5\}$.
- Used the raw counts in bins without scaling and transforming as explained in section 2.
- Traditional time series models, modeling bin counts as a function of time.

3.4 Summary of Coefficients and Prediction Errors

Tables 1,2 and 3 show the summary statistics of estimated coefficients and prediction errors for the models 1,2 and 3 respectively. After training the data on first 5000 observations, the three models were re-estimated 1120 times by adding one observation at a time. By observing the minimum, maximum, variance etc. of the coefficients, it is apparent that the coefficients are extremely stable in all the cases. The variance of all the coefficients is almost zero in all the three models. All the prediction errors have a mean close to zero.

Table 1

Summary Statistics of the Coefficients and Prediction Error of Model 1

Summary of the Estimated Coefficients, Observations = 1120						
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Variance
4.183e-01	4.186e-01	4.187e-01	4.187e-01	4.189e-01	4.190e-01	4.031952e-08
4.178e-01	4.179e-01	4.180e-01	4.180e-01	4.182e-01	4.184e-01	3.198208e-08
5.822e-01	5.825e-01	5.826e-01	5.826e-01	5.828e-01	5.829e-01	3.318152e-08
-6.248e-05	-6.077e-05	-5.910e-05	-5.825e-05	-5.767e-05	-5.116e-05	1.059726e-11
-2.166e-04	-2.145e-04	-2.135e-04	-2.135e-04	-2.128e-04	-2.101e-04	2.269010e-12
4.750e-04	4.766e-04	4.778e-04	4.810e-04	4.809e-04	4.960e-04	4.315627e-11
1.862e-03	1.866e-03	1.872e-03	1.873e-03	1.877e-03	1.890e-03	5.895899e-11
-2.700e-05	-1.860e-05	-1.556e-05	-1.757e-05	-1.450e-05	-1.314e-05	1.905561e-11
-1.691e-03	-1.674e-03	-1.671e-03	-1.674e-03	-1.671e-03	-1.667e-03	3.754893e-11
1.328e-03	1.425e-03	1.434e-03	1.429e-03	1.445e-03	1.460e-03	6.509778e-10
Summary of the Prediction Error						
-0.0044480	-0.0010500	-0.0002062	-0.0001817	0.0006204	0.0091600	1.842262e-06

Table 2

Summary Statistics of the Coefficients and Prediction Error of Model 2

Summary of the Estimated Coefficients, Observations = 1120						
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Variance
0.508800	0.509700	0.510300	0.510300	0.511300	0.511500	8.924762e-07
1.057000	1.057000	1.057000	1.057000	1.057000	1.057000	3.953955e-08
-0.002523	-0.002483	-0.002476	-0.002474	-0.002471	-0.002416	4.729647e-10
0.001550	0.001568	0.001637	0.001640	0.001697	0.001760	4.387264e-09
0.004925	0.004950	0.004967	0.004968	0.004989	0.005018	5.573793e-10
-0.014550	-0.014540	-0.014450	-0.014450	-0.014390	-0.014310	7.447432e-09
-0.001956	-0.001922	-0.001884	-0.001879	-0.001821	-0.001812	2.545509e-09
-0.024240	-0.024110	-0.024030	-0.024020	-0.023880	-0.023860	1.740845e-08
-0.014380	-0.014230	-0.013690	-0.013730	-0.013300	-0.013050	2.199717e-07
Summary of the Prediction Error						
-0.035340	-0.012930	-0.005522	-0.005200	0.001324	0.072120	0.0001425064

Table 3

Summary Statistics of the Coefficients and Prediction Error of Model 3

Summary of the Estimated Coefficients, Observations = 1120						
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Variance
0.698600	0.699700	0.700500	0.700600	0.701800	0.702100	1.447957e-06
1.038000	1.038000	1.038000	1.038000	1.038000	1.039000	5.637011e-08
-0.002363	-0.002325	-0.002301	-0.002306	-0.002291	-0.002248	6.568119e-10
0.002473	0.002494	0.002528	0.002542	0.002575	0.002658	3.063131e-09
0.004073	0.004103	0.004125	0.004124	0.004149	0.004179	7.981963e-10
-0.015750	-0.015730	-0.015650	-0.015650	-0.015590	-0.015520	6.212995e-09
-0.001749	-0.001724	-0.001684	-0.001686	-0.001645	-0.001632	1.608052e-09
-0.019060	-0.018930	-0.018830	-0.018840	-0.018700	-0.018670	1.756437e-08
-0.015900	-0.015580	-0.015100	-0.015130	-0.014690	-0.014440	2.331910e-07
Summary of the Prediction Error						
-0.039150	-0.013750	-0.006141	-0.005655	0.001224	0.062070	0.0001568749

4 Distribution Fitting of the MET 1 Data

4.1 Methodology

The counts in each bin (1-6) form a histogram of the data that was fit to several likely candidate probability distributions. The fitting was done using Matlab software because of its built in routines for various probability distribution functions (PDFs). The distributions fitted were:

- exponential
- beta
- gamma
- weibull
- lognormal

The histogram of the data, either for each sensor or for the combination of all five sensors, had the features of an exponential decay, with a long tail that actually began to increase for the last bin (a U shape). We assumed that this increase in counts for bin 6 was an artifact of the measurement process and that had higher bin resolutions been possible, the tail would continue on its decreasing trend. Therefore, it was not necessary to fit a distribution whose tail turned upwards. However, such a shape is possible with a beta distribution. Therefore, it was chosen.

Another feature of the histogram of the data was the behavior at 0 micron particle size and 0 counts (relative frequency). If one assumes that this is the true intercept, then unimodal distributions which begin at (0, 0) and rise sharply to capture the very high frequency counts in the first bin are potential candidates for fitting. The lognormal, weibull, gamma and beta can exhibit such behavior. It should be noted that the beta and weibull can also be shaped in a reverse J , like the exponential, such that the relative frequency count is infinity at particle size 0 microns.

Therefore, the beta was chosen for its flexibility in possible shapes: reverse J , U , and unimodal with intercept (0, 0). The lognormal was chosen for its potential as indicated from previous studies. The exponential was chosen to capture the basic decay shape of the data, reverse J . The weibull is a more general form of the exponential, making it a viable candidate. Finally, the gamma was chosen for its potential of capturing the tail behavior indicated by the histogram of the data.

While other distributions are certainly viable candidates, time restrictions prevented a more thorough investigation of these and of general distribution families such as the Pearson families or the exponential family (of which each of these chosen five is a member).

To test the goodness of the fits of these five distributions to the data, a Kolmogorov-Smirnov (KS) one-sample, two sided-test was proposed (as found in SPLUS). While other tests for comparing distributions are available (such as those based on Kullback-Liebler information or Jeffery's measure), time did not permit programming them for use.

4.2 Results

While fits were done for all five sensors, individually, for the five PDFs, as with the modeling efforts in the previous sections, only the results for the combined sensor data are listed here. Because of software limitations, the counts for all bins were divided by 100 for the fitting and plotting. Table 4 shows the parameter values for each of the PDF fits.

Table 4
Parameters of Fitted Distributions

Probability Distribution	Parameter(s)
beta	$\alpha = 3.122, \beta = 39.480$
lognormal	$\mu = 0.701, \sigma^2 = 0.123$
weibull	$\lambda = 1.394, \beta = 1.644$
gamma	$\alpha = 3.547, \beta = 0.205$
exponential	$\lambda = 1.376$

Mathematica was used to plot the distribution fits along with the data fits. Figures 4a-e show each of the five PDFs as dashed curves with the MET1 data as a solid curve. The particle sizes (microns) were normalized (to fall between 0 and 1) for fitting and plotting the beta distribution, to accommodate the fact that the two parameter beta ranges from 0 to 1. This rescaling is a disadvantage in using the beta distribution.

Visually the lognormal peaks too high and tails off too slowly for the data. While the weibull peaks too low. The exponential decay is not quite in alignment with the data's decay, but it is good at capturing the lower tail. The gamma also produces a reasonable fit all around, as does the beta. The beta does the best job capturing the tail, while also capturing the decay and the peak.

Kolmogorov-Smirnov tests for goodness of fit were run to determine if any of the five distribution fits, statistically matched the total particle size counts. Using a 5% significance level for the two-sided test, none of these distribution fits matched the data.

5 Future Studies

To fully test the models from section 3 and determine their viability for discrimination of releases, release data is required. At this point we can only speculate that setting $\mathbf{y} = 0$ will produce different coefficients from those found in models where $\mathbf{y} = 1$. Nonetheless, the models are available for such an exercise should release data be available in future studies.

UV background data was made available for similar analyses. However, time and money for the project ran out before models and/or distributions could be fit to that data. Again, the ultimate test of such models would be testing them against release data.